Roundtable
HIGHLIGHTS
March 2019

Roundtable on Data Science Postsecondary Education

Meeting #9 - December 10, 2018

The ninth Roundtable on Data Science Postsecondary Education was held on December 10, 2018, at the Keck Center of the National Academies in Washington, DC. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to learn about academic, government, nonprofit, and private sector projects promoting data science for socially desirable outcomes and their intersection with education and hiring; and to explore how socially motivated projects and topics can engage and excite students. This Roundtable Highlights summarizes the presentations and discussions that took place during the meeting. The opinions presented are those of the individual participants and do not necessarily reflect the views of the National Academies or the sponsors. Watch meeting videos or download presentations at **nas.edu/DSERT**.

Welcoming Roundtable participants, Kathleen McKeown, Columbia University, commented that many students in data science, computer science, and statistics courses are eager to "give back" to their communities through the practice of data science for social good. She highlighted ethical concerns raised during previous meetings of the Roundtable, such as potential bias in machine learning and fair artificial intelligence (AI), which are important to revisit in discussions of how data could be used for social impact.

AN INFORMAL DISCUSSION ABOUT DATA SCIENCE FOR SOCIAL GOOD

D.J. Patil, Devoted Health and former Chief Data Scientist, White House Office of Science and Technology Policy

Patil explained that the Chief Data Scientist for the United States, a role that is currently vacant, works to ensure that data are used responsibly to benefit all people uniformly instead of to divide or oppress individuals and communities. As data officer positions have been created within the federal government, at state and city levels across the country, and throughout the world, Patil is hopeful that the current administration will find a way to leverage this role.

While many researchers are focused on AI and algorithmic bias, Patil noted that data collection, use, safety, and security, as well as appropriate policy making around data, are basic concepts worthy of increased attention. *Ethics and Data Science* (written by Patil, Hilary Mason, and Mike Loukides) identifies ethical constructs lacking in organizations, such as a dissent channel, a checklist for product launches, and standard principles for ethical data use. He suggested that ethics and security be integrated throughout data science curricula and that future data scientists receive increased liberal arts training. He championed the role of 2-year institutions in offering introductions to data science for social good, and he advocated for Congress to support free education from 2-year institutions for all Americans.

Patil identified ways in which data science could be used to benefit society. For example, various technologies could have been used during Hurricane Katrina to predict how many people would evacuate and from which

areas, to detect where bridges were washed away, to locate people sheltering on rooftops, and to direct boats engaged in search and rescue missions. Data science could also be used to help police departments compare data across state lines, as no infrastructure currently exists to do so. However, Patil emphasized that transparency remains an issue, especially for applications in the criminal justice system. The mental health space is already benefitting from data science applications with the development of a crisis text line to help meet the demand of mental health emergencies. Patil added that data science and AI could have a substantial impact on basic logistics and transportation problems. He emphasized that one does not need access to a large data set to impact society and suggested contacting local food banks or shelters to find out if their challenges could be addressed with data science.

Uri Treisman, University of Texas, Austin, observed that when government agencies fail to manage crises, citizens often organize responses. However, getting data quickly and optimizing resources remains a challenge; local volunteers need to be trained to use data science in emergency situations. Patil agreed that these "digital humanitarians" need guidance on how best to create infrastructure and coordinate so as to be most effective. Jessica Utts, University of California, Irvine, asked Patil for advice on structuring a data ethics course. Patil encouraged faculty to integrate ethics throughout the curriculum—referencing Prof. Ed Felten's, Princeton University, case study approach as a model—instead of offering only one course on ethics and security. He directed participants to view and contribute to a collection of curricula from faculty across the country. Mehran Sahami, Stanford University, asked Patil to talk more about the importance of liberal arts education and the most useful tools for data science. Patil described liberal arts' emphasis on formalism, creativity, and framework development as invaluable in preparing to solve industry and societal problems.

FROM CLASSROOM TO CLINIC: DATA SCIENCE FOR SOCIAL GOOD FELLOWSHIPS AND THE LESSONS DATA SCIENCE EDUCATORS CAN LEARN FROM THE MEDICAL PROFESSION

Matt Gee, University of Chicago and BrightHive

Gee described the Data Science for Social Good (DSSG) program as an immersive fellowship in which aspiring data scientists learn how to map data methods and tools to social problems in partnership with a government agency or nonprofit organization. Gee said that DSSG builds a community of open, ethical, collaborative data science practice through research and development, lectures, workshops, and events. In its first year, DSSG received more than 600 applications but chose only 36 fellows to participate in the program. DSSG looked for partner organizations with important problems, leadership buy-in, access to data, staff capacity to work with data, and a commitment to implementing solutions. After defining goals, determining what actions would be taken, identifying what data were available internally and what data would be needed, deciding what analysis needed to be done and how it would be validated, 14 projects emerged and the first cohort of fellows arrived in May 2013. In working with both their partner organizations and their DSSG mentors, fellows learned to consider the social and ethical implications of data used in decision making as well as strategies to communicate with diverse audiences. Project outcomes have included solutions to predict heart attacks, to anticipate school drop-out rates and improve graduation rates, to help state governments save money on energy bills, and to help aid organizations respond to crises faster. Partner organizations emerged with an understanding of how to view data as an asset, Gee said, while fellows learned that data science tools, when used responsibly, may amplify one's ability to do good. During its 6 years, DSSG has engaged more than 224 fellows from all over the world in 70 projects.

Although the program has made great progress, Gee explained that the title "Data Science for Social Good" implies a moral superiority for data science that helps nonprofits and government agencies, and reduces data science for social good to something one does in his/her spare time. He emphasized that all data science should be grounded in a sense of the good; instead of *doing data science for good*, professionals should continually *do good data science*. As educators consider the future of data science training, Gee suggested turning to long-established professions, such as medicine, and learning from their experiences. He referenced Paul Starr's *The Social Transformation of American Medicine* in his rationale for new data science pedagogy. First, he explained that because data science has gained popularity, economic power, and cultural cache quickly, data scientists are often unaware of the potential consequences of their work. Data science education is currently failing in that it is taught at a distance, with clean data sets separated from social context. Instead, Gee continued, students need to be taught about developing personal accountability and avoiding algorithmic tyranny, in

which algorithms lead rather than inform decision making. For example, DSSG fellows spend the first 2 weeks of the program working *without data*, talking with project partners, and gathering context. Second, he explained that data science would benefit from the development of professional norms—for instance, choosing service over profit when the two conflict, so that consumers know their best interests are considered when working with their data. Gee referenced The Global Data Ethics Project as an example of the profession's attempt to adopt ethical principles. Third, he commented that it is important for the data science profession to attract and retain the best and brightest minds. He noted that 75 percent of adults under age 35 are willing to take a pay cut to do social good work.

Moving forward, postsecondary educators could add clinical practice requirements to data science programs. Although this could be both complicated and expensive, Gee commented that this would allow students to explore the social context of where data are generated and will be used, developing the analogue to medicine's "bedside manner" for data science. Educators could add written and verbal discussions of the social and ethical implications of data sets and models into problem sets in data science coursework instead of relegating ethical conversations to a single course. Educators could also provide guidance to employers for incorporating ethics case studies into hiring, apprenticeship, and mentorship opportunities. Taking these steps to improve data science training, Gee said, could render data science as more of a "healing profession with deep purpose and moral authority." Michael Pearson, Mathematical Association of America, asked Gee if DSSG includes discussions of how data science will inform policy or hold policy makers accountable for data misuse. Gee acknowledged that the program would benefit from more discussions about "data misuse" as well as "data missed use."

TEACHING DATA THAT MATTERS

Rahul Bhargava, Massachusetts Institute of Technology Media Lab

Bhargava began with a moment of silence to acknowledge that many people still face discrimination working in the space of data science and to honor the history of Title IX, which has provided instruments to help address this problem. In discussing the concept of data storytelling, Bhargava noted that how information is presented to an audience impacts how it will be understood—viewers are often distanced from the lived reality of data. He described the separation that exists between the desire to do something valuable with data and the respect for the experience of the person represented by the data. This notion of respect is accompanied by a question of responsibility: is an algorithm designer responsible for what happens to an algorithm user?

Bhargava explained that powerful people have used data to subjugate those without power throughout history. For instance, Egyptian leaders created a census to catalogue laborers for the construction of pyramids. This history has to be acknowledged and challenged by those who wish to use data for good, he continued. Both historic and contemporary counter-efforts exist: Predictive models were developed in the 17th century to prevent the Bubonic Plague, and W.E.B. DuBois used infographics to catalogue and share the life experiences of former slaves. Currently, the Data for Black Lives organization works to eliminate the presence of bias in data. In all of these cases, data were used to tell alternate stories about matters of social importance. Teaching "data that matters" presents an opportunity for students to better use and understand real data, to ask hard questions and take risks, and to balance learning objectives with personal interests. Bhargava teaches a cross-disciplinary course, hosted by the MIT Humanities department, called Data Storytelling Studio, in which students "consider the emotional, aesthetic, and practical effects of different [data] presentation methods." This course is offered via MIT's Open Courseware.

He described three student projects from this course in which data sets were put into context to inform actions: (1) a board game, comprised of refugee data, that people "play" at a fundraiser to better understand the refugee experience and hopefully donate to the cause; (2) an inverted map of real stop-and-frisk data, accompanied by a satirical data journalism story; and (3) a data-driven game, based on Food and Drug Administration data, to teach children about the roles that bees play in the environment. Bhargava said that his classroom is a "playground" where students "flex their data muscles" in a safe learning space. So that other educators can access hands-on data-storytelling activities, this open source content is available through the Data Culture Project. Alfred Hero, University of Michigan, wondered how Bhargava achieves a convergence between his course and more traditional data science methodology courses since many students enrolled in the latter may not enroll in the former. Bhargava said that he recruits

students for his 30-person course; those students then advertise the course in their departments.

OPEN DISCUSSION

Program Development

Nicholas Horton, Amherst College, noted that DSSG serves as a model of integrated co-curricular experiences, but he wondered about the barriers to rolling out similar programs at less-well-resourced institutions. Gee said that while challenges vary by institution, few institutions offer a clear home for such a program or the faculty and budget lines to support it. He emphasized the value of creating a dedicated co-curricular space. Bhargava noted that MIT's Media Lab, known for its anti-disciplinarity, has positioned itself at the intersection of numerous fields, is well supported, and attracts great students and faculty. He noted that no single recipe for success exists for all institutions. Bill Howe, University of Washington, wondered if emphasizing the liberal arts and injecting more social context into data science programs could cause some students to lose interest in the courses, either because they are different than they imagined or because they require messy project work. Bhargava said that truths about fields are always evolving; faculty should help students reset their assumptions and build a new knowledge base. He addresses similar student concerns through team design, pairing students with different perspectives, learning goals, and work habits. Gee said that some fellows consider leaving the program each year because they dislike the amount of time spent talking with project partners or navigating team politics; however, most ultimately realize that this "messiness" *is* the benefit of doing clinical practice.

Community Partnerships

Deb Agarwal, Lawrence Berkeley National Laboratory, commented that although short-term problem-solving engagements have some value, she asked whether DSSG fellows have the opportunity to study and learn from the efforts of previous cohorts. Gee responded that the fellows discuss why projects were not chosen, which helps them understand "messy" issues they may face; however, he noted that he might implement Agarwal's idea with a future cohort. Rachel Levy, Mathematical Association of America, wondered if DSSG project partners have the capacity to test and use the solution provided by the fellows and have the independence to modify it. She emphasized the value of thinking about tools as opportunities not only for the fellows but also for the project partners. Gee described three possible considerations to build better capacity within the partner organizations: (1) right-size the project to the course and the timeline; (2) provide cross-semester or cross-year continuity for a project; and (3) ensure that training for the partner is built into the curriculum. Bhargava mentioned that he no longer develops community partnerships in his course because one semester is insufficient to cultivate such relationships.

Ethical Considerations

Sahami said that many computer science faculty are uncomfortable teaching ethics because they lack the relevant training. While a philosophy department could offer a multi-disciplinary course, he wondered what other strategies could be used to teach ethics in a meaningful, balanced way. Gee and Bhargava suggested that simply exposing students to the appropriate set of questions, without necessarily providing the foundational text, helps prepare them to continue to learn on their own. Treisman said that because of the power and potential of data science, a rich liberal arts background should be embedded in data science education. Students have to understand how to enter into the social worlds in which they are going to use data if the objective is to empower people, he continued. Treisman emphasized that all academic departments, not just philosophy, have an obligation to attend to the social, ethical, and moral development of students. Bhargava supported the notions of learning data in practice and challenging arbitrary disciplinary boundaries. Jeffrey Ullman, Stanford University, questioned whether educators and researchers have the right slant on the matter of data consent, as the notion of data privacy is a modern construct. He said that because Google and Facebook are free platforms, consent is a difficult concept; if the companies were to charge users to opt out of data collection, lower-income users lose access to privacy protection. He emphasized the need to think carefully about allowing data consent, pointing to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) as an example of how a codification of privacy rights can have unintended consequences (e.g., in HIPAA's case, complicating patients' ability to communicate with medical professionals). Gee said that challenges arise when DSSG project partners have ill-defined data privacy policies. An audience participant wondered about Gee's previous analogy of data science to the field of medicine: many more stakeholders exist in data science than medicine, and negative consequences outside of the client relationship are possible. Gee agreed that the analogy between data science and medicine is imperfect;

however, he noted that data science faces very similar questions about "professional sovereignty."

DATA, DESIGN, AND ENGAGEMENT: LESSONS FROM 30+ DATA SCIENCE FOR SOCIAL GOOD PROJECTS

Peter Bull, DrivenData

DrivenData has worked on more than 50 projects with nonprofits, social enterprises, and corporate social responsibility groups, Bull explained, and it tries to figure out how to solve organizations' problems with machine learning or data science tools, using the data assets that they already have. An organization's problem is posted online, and a community of data scientists proposes algorithms to solve it. DrivenData selects the best-performing algorithm and assists the organization with implementation. DrivenData has run more than 30 competitions during the past 5 years, with participation from a community of more than 35,000 data scientists from across the world.

Bull described three example projects. The first project helped a school district approach budget benchmarking in the absence of structured data about school spending. DrivenData helped build an algorithm for the school district to generate predictions for spending as well as information about what part of the budget was being used. This automated process replaced the approximately 300 staff hours per year that were spent analyzing spreadsheets with similar information. The second project helped a community improve its strategy for capturing water from coastal fog with mesh nets, DrivenData used data from weather stations located next to these mesh nets to try to predict their yield. This work prompted the community to prioritize the placement of new fog nets. The third project helped to prioritize health inspections for Boston restaurants using data from 4 years of health code violations combined with Yelp reviews and ratings. With this new method in place, inspectors were able to find 25 percent more violations and thus better protect citizens.

Bull explained that achieving the highest accuracy is not always the desired outcome when building a model. Instead, the desired outcome is how the accuracy works in concert with other goals and metrics for success. With this in mind, DrivenData hosted a new type of competition, Concept to Clinic, in which contributors earn points and achieve visibility by submitting their work to an open source repository. This adds an element of collaboration to the competition and promotes sharing throughout the process instead of upon completion, Bull continued. He described DrivenData's other open source projects, including Cookiecutter Data Science, a standardized project structure for doing data science work, and Deon, an ethics-checklist generator for projects. DrivenData also engages directly with organizations to solve data science problems. In closing, Bull shared a Data Impact Field Guide, with concrete challenges to consider before engaging in a project:

- *Finding a project*. Bull said that this is the most difficult part of the process and where the greatest need exists. Ninety-five percent of the time, organizations want help measuring impact. However, data scientists may not be the best equipped to do this in a short amount of time. If one thinks about impact measurement as early as during the data collection stage, the majority of the work will be done by a domain expert, whereas if one thinks about impact measurement during data analysis, the majority of the work will be done by a data scientist.
- *Launching a project*. Bull noted that because social sector organizations exist for the public good, they demand higher attention to data ethics. For example, questions about security, explainability, and responsibility arise during the data collection, modeling, and deployment phases, respectively. He asserted that better ethics develop through increased practice.
- *Running a project*. A project should build trust and empathy between the user and the technologies by embedding ideas from human-centered design thinking into the data science process. A human-centered data scientist will go to the field and observe data being generated; design plans with the user by iterating jointly on prototypes; assess outcomes both quantitatively and qualitatively; and be honest about and learn from failures.
- *Wrapping up a project.* The capacity gap between the social sector and either industry or academia is wide and can jeopardize solution hand-offs. There is also a shortage of more than 140,000 data scientists in industry, a problem felt heavily in the social sector.

TEACHING PEOPLE TO THINK WITH DATA

James Hodson, AI for Good Foundation

Hodson explained that the AI for Good Foundation was established in 2014 after a series of workshops at Stanford University about the status of AI and future

innovation. Participants discussed core problems, breakthrough methodologies, and social impacts. After the workshops, he continued, it became clear that a bridge between research laboratories and government, industry, and nonprofit stakeholders was needed. Questions emerged about how AI aligns with the notion of social good as well as how communities could be built to enable long-term change. In response, the AI for Good Foundation adopted the United Nations' 17 Sustainable Development Goals as its framework. Although these goals are unlikely to be attained in the near term, Hodson noted, they raise questions about how to solve this generation's challenges. The AI for Good Foundation continues to build the capacity to reach these goals through partnerships with academic laboratories. He said that cross-departmental initiatives at academic institutions, in combination with engagement from actors on the ground, are promising.

He presented a potential definition of data science: a set of algorithmic methods and engineering practices that need a channel for development and adoption within empirical research. He added that industry and society need data literacy to harness the value of data and to aid in solutions to a wide variety of problems. Hodson said that it is important for students to understand the realities of the challenges people are facing in the real world. He emphasized that data science does not need to be housed in a standalone department because it should not be viewed as a different field. He explained that each academic discipline has its own long-established tradition of working with data, and, although it would require additional faculty training, each discipline could teach important aspects of data science within its department. Academic institutions have a responsibility to train people to go into industry and government to solve hard problems with data, rather than training everyone to be a data scientist, he continued.

Hodson said that society should embrace data-driven science; data literacy across campus; cross-disciplinary research and teaching resources; open infrastructure, data, and methods; data innovation hubs; data science for social good; and diversity. The main barriers to achieving these goals are that the methods are often taught independently from the research process; students are seldom taught how to evaluate, clean, and merge data; and the teaching of applied data science in a laboratory setting is too short, too stylized, and has no impact. Hodson noted that discussions about ethics should not be motivated only by regulatory purposes. To truly bring social impact into the data science classroom, one semester of instruction is insufficient, he continued. Sahami asked to what extent students should be engaging in projects with real social impact and measuring results versus understanding the issues and methodology. He noted that, in academia, faculty are often constrained by time, expertise, and resources. Hodson agreed that merging best educational practices with social impact is challenging. While he acknowledged that there is an opportunity to use projects as gateways for continuing interaction, he said that they are not necessary to teach the fundamental principles of data science.

CAN AI REDUCE GANG VIOLENCE OR CAUSE MORE HARM?

Desmond Patton, Columbia University

Patton's current work uses qualitative methods, machine learning, and community expertise to better understand how social media provides a window into gang violence. SafeLab's interdisciplinary team of social scientists, computer scientists, and domain experts develops technology tools to support the prevention of gang violence. Patton was motivated to study this area by the rise of crime in Chicago—764 homicides occurred in 2016, most of which involved guns, public spaces, and prior altercations, many of which were described in social media posts.

SafeLab studied how a now-deceased gang member, Gakirah, narrated her life on social media and how other people responded to her posts. Many of the posts were difficult to understand in terms of language, context, and nuance, so a methodological approach was needed to understand the data. During the first stage of the contextual analysis, the research question and study population were clarified, the social media corpus was created, and domain experts (i.e., gang members and other youth in the community) were identified. After annotators received training, they began to code the data and to develop a baseline interpretation. Annotators then created descriptions informed by the context of the social media post, and machine learning was used to label data as "loss," "aggression," or "other." Domain experts would then review the labels and help reconcile the interpretations by providing additional context. The labeled data sets were fed into natural language processing algorithms, which developed additional tags and labels to translate the social media data into standard English.

Patton's team is developing greater accuracy and leveraging more context; in 2018, it developed a new labeled data set, six times larger than the previous, and integrated neural net approaches. The team has also established new partnerships with computer vision specialists so that information can also be collected from images, which tend to better identify aggression and substance abuse, according to Patton. He explained that ethics is especially important in this line of work: the team is careful in how it uses information about aggression in young men and women of color, has its annotators sign non-disclosure agreements, and refrains from sharing publicly any images from the data set. This work provokes a conversation about the importance of data in context—Patton's team is developing a conceptual framework to theorize how social media policing can negatively impact communities of color and is creating digital interventions for youth. Ultimately, Patton's goal is to build empathy and to drive behavioral change, as young people may not understand the consequences of their digital footprints.

Eric Kolaczyk, Boston University, asked about the challenges and lessons learned during annotator training. Patton responded that a main challenge was trying to figure out how to best support the diverse annotators. He noted that he did not anticipate the way that "life would get in the way" for the young people serving as domain experts and added that challenges exist in maintaining relationships with them. The social work students had to learn how to treat the user as a whole person and how to interpret more accurately. Patton also cited a need to be aware of the triggers that can happen for annotators confronted with disturbing posts—for example, about violence toward women. In response to McKeown, a member of Patton's research team, Patton commented that the social work and computer science students worked well together and pushed each other toward the best solutions. The social work students taught the computer science students strategies to confront real-world problems and challenges, while the computer science students taught the social work students to develop data literacy and to ask the right research questions. In response to a question from Ullman about law enforcement's use of electronic footprint monitoring, Patton suggested that people should challenge and critique the methodology as well as understand the context. He emphasized the importance of using these techniques equitably and applying them across demographic groups uniformly. Louis Gross, University of Tennessee, described a workshop he will host in May 2019 on the mathematics of gun violence and potential impacts for alternate interventions, and Patton encouraged him to include social scientists in the conversation to think about data patterns.

OPEN DISCUSSION

Data Science Education

Ullman asked the speakers to outline the technical differences between "data science for social good" and "data science." Bull responded that it is important to give all data scientists a concrete process to ask the right questions in order to understand the domain they are working in, especially when it comes time to hand off a solution to an organization. Hodson said that there is great opportunity to make social impact through data science, but that is not the most important part of rethinking data science education. He reiterated that data science is a set of processes and methodologies in which all departments should partake as opposed to a separate discipline. He encouraged cross-departmental collaboration instead of teaching data science as an isolated subject. Bull said that data science may face similar challenges to the field of software engineering in finding a disciplinary home that has both a particular set of skills and domain-specific research questions. Hero asked how data science will scale to meet high student demand if it is not housed in a separate department. Hodson said that courses that are not necessarily department-specific will have to be created.

Collaboration and Management

Kolaczyk asked how people in academia could best interact with Bull's and Hodson's organizations. Hodson said that the AI for Good Foundation has done many programs jointly with universities and public research institutes (e.g., workshop series, co-teaching). He noted that the foundation tries to unite researchers, students, and community stakeholder groups; it helps external organizations understand where they need advice and interaction and helps researchers understand how theoretical research can be applied. Bull suggested that academic institutions avoid partnering with DrivenData because doing so could create a bottleneck; however, the lessons learned from working with organizations could be used as resources for educators who wish to set up their own projects. Educators could set up long-term partnerships with other organizations across multiple years and multiple cohorts of students, Bull said. Treisman noted that the relational trust needed when working in political environments is more complex than when trying to help a business optimize sales,

for example. He highlighted cycles of interaction between data users and data owners, in which trust has to be built around everyone knowing and following the same rules. He wondered if anyone has written descriptions of these processes as well as how we might make it easier for people to learn how to do this work. Bull agreed that working with social-sector organizations is often more difficult because their metrics of success are undefined and that building trust is critical. He suggested that the data science community think carefully about the best way to engage with these organizations. Hodson noted that organizational behavior research may provide insight into these areas. He said that the structure of the institutions that people are working within have to change to allow for these new types of interactions. Treisman added that the role of design expertise is often underestimated when organizations attempt to improve. He explained that data management/optimization techniques, institutional mechanisms for knowledge management, and clever design are essential, most of which does not come from technical, mathematical tools. Bull appreciated Treisman's description and agreed that DrivenData faces a challenge of balancing creativity and knowledge management with technical know-how. Hodson agreed but added that educators have a responsibility to teach people how to develop architectures that will lead to better outcomes.

SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Following the presentations and open discussions, Roundtable participants divided into three groups to discuss specific themes from the day. The first group discussed integrating data context into students' coursework. On behalf of her group, Levy explained that each discipline has a different way of facilitating communication between domain experts and technical experts. It is important for students to develop an appreciation for what each person contributes to such a conversation. While it may be possible to teach students how to have those conversations, Levy continued, it is a skill that needs to be practiced and developed over time. She said that students should explore and experience misunderstandings of language, culture, biases, assumptions, and constraints in order to be better practitioners in context. Levy's group also questioned the use of the phrase "social good," as its meaning may vary by context. Her group said that conversations about what "social good" means, who defines it, and who benefits from it should be included in data science curricula.

The second group discussed the benefits and drawbacks of increased training around data science for social good. On behalf of his group, Ullman acknowledged that some students and faculty are only interested in the theory of a subject rather than its practical application. He used mathematics as an example of a discipline that has been driven by theory, successfully, for 3,000 years. In data science, he continued, people who are interested in developing new machine learning models without paying attention to what they will be used for could create problems. He suggested that it may be ineffective to orient data science education programs toward people who are uninterested in how their ideas will be applied. When people are forced to work in diverse teams (e.g., data scientists and domain experts), people step outside of their comfort zones and explore broader issues. Ullman's group advocated for a curriculum with a solid mix of theory and practice and noted that a flipped classroom is one way to facilitate such a curriculum.

The third group discussed how to incorporate ethics in a responsible and informed manner across the curriculum. On behalf of his group, Sahami explained that definitions of "social good" and "ethics" remain unclear. He suggested integrating ethics into data science instead of discussing it as a separate entity so as to better develop ethical behavior. Although there are many layers in the technology stack—for example, who is responsible for how technology is used—issues of ethics, social justice, and societal good are often combined and thus not considered adequately. Sahami noted that data science education and practice could benefit from the best ethical practices of other more established communities and that alternative models could be embedded across multiple disciplines. Sahami's group also discussed the potential for those in leadership to speak more openly about issues of ethics so as to make the concept more accessible to young people. Sahami pointed out that data science does not yet view itself as a profession like medicine, which has a clear code of ethics. Utts added that faculty are trained with integrity in their disciplines and should pass those principles along to their students in every course, which Howe connected to Gee's earlier discussion of "professional sovereignty." Kolaczyk wondered if society has reached a point where the potential to do good or harm is at a completely different scale than ever before, forcing practitioners and educators to wrestle with larger issues. Treisman noted that the data science community can influence the infrastructures that currently stipulate ethical behavior.

ABOUT THE ROUNDTABLE: The Roundtable on Data Science Postsecondary Education is supported by the Gordon and Betty Moore Foundation, the National Institutes of Health, the National Academy of Sciences W. K. Kellogg Foundation Fund, the Association for Computing Machinery, the American Statistical Association, and the Mathematical Association of America. Within the National Academies, this roundtable is organized by the Committee on Applied and Theoretical Statistics in conjunction with the Board on Mathematical Sciences and Analytics, the Computer Science and Telecommunications Board, and the Board on Science Education. Roundtable meetings take place approximately four times per year. Please address any questions or comments to Ben Wender at **bwender@nas.edu**.

ROUNDTABLE MEMBERS PRESENT: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Lise Getoor, University of California, Santa Cruz; Alfred Hero III, University of Michigan; Nicholas Horton (via webcast), Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Rachel Levy, Mathematical Association of America; Nina Mishra, Amazon; Michael Pearson, Mathematical Association of America; Mehran Sahami, Stanford University; Uri Treisman, University of Texas, Austin; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

GUESTS PRESENT: Tensae Andargachew, New Jersey Institute of Technology; James Angelo, Leidos; Ted Avraham, The Jewish Student Satellite Initiative; Rahul Bhargava, Massachusetts Institute of Technology Media Lab; Cheri Borsky; Peter Bull, DrivenData; Michael P. Cohen, American Institutes for Research; Richard Esposito, Bureau of Labor Statistics; Adam Fagen, BioQUEST Curriculum Consortium; Matt Gee, University of Chicago and BrightHive; Lauri Goldkind, Fordham University; Louis Gross, University of Tennessee; Doug Hague, University of North Carolina, Charlotte; Robert Hershey, Robert L. Hershey, P.E.; James Hodson, AI for Good Foundation; Kristin Jenkins, BioQUEST; Benjamin Kallen, Lewis-Burke Associates; Brian Kotz, Montgomery College; Kathryn Kozak, American Mathematical Association of Two-Year Colleges; Zenobia Liendo, George Washington University/University of California, Berkeley; Elizabeth McDaniel, Institute for Defense Analyses; John McNutt, University of Delaware; Sharon McPherson, National Science Foundation; Peter Mecca, George Mason High School; D.J. Patil (via webcast), Devoted Health; Desmond Patton (via webcast); Columbia University; John Rowan, Coding Dojo; Frank Sanacory, The State University of New York College at Old Westbury; Yla Tausczik, University of Maryland iSchool; Jeremy Wojdak, Radford University/QUBES; Brian Wright, George Washington University; Li-chiung Yang; Maryam Zaringhalam, National Library of Medicine.

NATIONAL ACADEMIES' STAFF PRESENT: Shenae Bradley, Linda Casola, Michelle Schwalbe, and Ben Wender.

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org